Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks

Heidemann Andersen, Asger; Haan, Jan Mark De; Tan, Zheng-Hua; Jensen, Jesper

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2018.2847459

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Non-Intrusive Speech Intelligibility Prediction using Convolutional Neural Networks

Asger Heidemann Andersen, Jan Mark de Haan, Zheng-Hua Tan, and Jesper Jensen

***Abstract*—Speech Intelligibility Prediction (SIP) algorithms are becoming popular tools within the development and operation of speech processing devices and algorithms. However, many SIP algorithms require knowledge of the underlying clean speech; a signal that is often not available in real-world applications. This has led to increased interest in *non-intrusive* SIP algorithms, which do not require clean speech to make predictions. In this paper we investigate the use of Convolutional Neural Networks (CNNs) for non-intrusive SIP. To do so, we utilize a CNN architecture that shows similarities to existing SIP algorithms, in terms of computational structure, and which allows for easy and meaningful visualization and interpretation of trained weights. We evaluate this architecture using a large dataset obtained by combining datasets from the literature. The proposed method shows high prediction performance when compared with four existing intrusive and non-intrusive SIP algorithms. This demonstrates the potential of deep learning for speech intelligibility prediction.**

***Index Terms*—non-intrusive speech intelligibility prediction, convolutional neural networks.**

I. Introduction

ALGORITHMS for Speech Intelligibility Prediction (SIP) attempt to predict the intelligibility of noisy or otherwise degraded recordings of speech as perceived by a group of average normal-hearing listeners. In this context, speech intelligibility is typically defined as the average fraction of words (measured in percent) that listeners can correctly understand in a given listening condition (i.e. for a given type of degradation). The use of SIP algorithms can be an advantageous alternative to carrying out time consuming and expensive listening experiments involving many test subjects. Such algorithms were first studied in the telephone industry, with the aim of quantifying intelligibility of speech transmitted via telephone, without relying on listening experiments [1], [2]. This research resulted in the Articulation Index (AI), which is an objective scoring of intelligibility, between zero and one [1]. This scoring is computed as a weighted sum of contributions from a range of non-overlapping frequency bands [1]. The band-wise contributions are, in turn, based on the long-term Signal to Noise Ratio (SNR) within each band [1]. The AI has been shown to have a nearly monotonic relationship with the measured intelligibility of speech masked by stationary noise. It has, furthermore, been shown that the AI can be interpreted as an estimate of the information capacity of a noisy communication channel [3]. An updated and ANSI-standardized version of the AI is known as the Speech Intelligibility Index (SII) [4].

Since the introduction of the AI, a considerable amount of research has been carried out within the SIP community. This more recent work often aims to provide predictions in conditions where the AI and SII fall short. These include conditions with fluctuating interferers, e.g. [5], [6], [7], [8], [9], [10], [11], conditions where the speech signal has been non-linearly degraded or processed, e.g. [12], [13], [14], [15], and conditions with binaural listening, e.g. [16], [17], [8], [18], [9], [19], [20], [21]. Several works have, furthermore, proposed SIP algorithms that account for hearing loss, e.g. [4], [22], [23], [24], [25]. On the theoretical level, efforts have been made to design algorithms that accurately model the auditory system in a physiological sense [26], [27], [10], [21], [24], [28], or which account for speech intelligibility from an information theoretical viewpoint [29], [30], [31], [32]. SIP algorithms can be classified according to the input signals required for making predictions. The AI and SII require access to clean speech and noise in separation (and therefore do not handle non-linearly degraded noisy speech) [1], [33], [34]. A second class of SIP algorithms makes predictions based on the degraded speech signal (i.e. that for which intelligibility is predicted), and either the clean speech in separation, e.g [12], [13], [24], [32], [11], or the clean noise in separation, e.g. [27], [10], [21]. Together, SIP algorithms that require knowledge of the clean speech signal or the clean noise signal are known as *intrusive* SIP algorithms [25].

Recently, research has been increasingly directed towards *non-intrusive* SIP algorithms, which predict intelligibility using only the degraded speech signal, e.g. [35], [22], [36], [37], [23], [38], [25], [39], [40], [41]. Such algorithms could be valuable for online assessment of speech intelligibility in signal processing devices such as hearing aids, or for other real-world applications where a clean signal cannot be obtained [25], [39], [41], [40]. A widely used non-intrusive SIP algorithm is the Speech to Reverberation Modulation energy Ratio (SRMR) [35], [22], [37], [23], [38], which has been developed to asses both intelligibility and quality of reverberant speech. It does so by quantifying the fraction of the input signal energy which can be attributed to low frequency modulations, based on the observation that speech signals are mainly characterized by low-frequency modulations [35].

The Short-Time Objective Intelligibility (STOI) measure is an intrusive SIP algorithm that has gained considerable popularity in the signal processing community [13]. This algorithm

Asger Heidemann Andersen is with Oticon A/S (e-mail: aand@oticon.com).

Jan Mark de Haan is with Oticon A/S.

Zheng-Hua Tan is with Aalborg University.

Jesper Jensen is with Oticon A/S and Aalborg University.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

predicts intelligibility from clean and degraded speech by averaging the sample correlation coefficient of short segments of clean and degraded envelopes across 15 one-third octave bands. The STOI measure has been shown to correlate well with measured intelligibility in conditions including different additive noise sources [13], noise reduction processing [42], [13], hearing-aid and cochlear implant processing [25], and noisy speech transmitted via telephone [43]. Later work has shown the STOI measure to be closely related to an estimate of information transmission [32]. The STOI measure has been extended to make binaural predictions [44], [45], as well as to make non-intrusive [40], or partly non-intrusive [39], [41], predictions. An extension of the STOI measure, which aims to increase prediction accuracy for speech masked by fluctuating noise, is proposed in [11]. Another extension aims to increase prediction accuracy by weighting contributions of speech segments according to the information content of the speech [46]. A noise reduction algorithm, based on maximizing the STOI measure, is presented in [47].

The SIP algorithms discussed until this point are typically based on heuristics, as well as simple models of the human auditory system, and in some cases information theory. The algorithms gain trustworthiness primarily from repeated displays of accurate prediction performance across different conditions. However, a number of works have attempted to partly or fully predict intelligibility using data-driven methods. One approach in this direction has been to use an Automatic Speech Recognition (ASR) system to transcribe degraded sentences, using the error rate as a measure of intelligibility [48], [49]. Another data-driven approach has been to non-intrusively estimate the output of an intrusive SIP algorithm [50], [51], [52], [53]. Both of these methods have the advantage that they do not rely on the availability of databases of measured intelligibility for training. Somewhat surprisingly, only limited attempts have been made to predict intelligibility directly from corresponding degraded speech signals [54], [55], [56], [57]. The recent success of deep learning, within both ASR [58] and speech enhancement [59], [60], suggests that such approaches could be successful within SIP. The absence of such work may be primarily due to the lack of widely available datasets of degraded speech and corresponding measured intelligibility.

In this paper, we use Convolutional Neural Networks (CNNs) to non-intrusively predict the intelligibility of degraded speech. We consider only monaural/diotic signals, thus avoiding the need to model binaural advantage. We consider speech signals which have been degraded by the addition of noise and by non-linear processing, making the proposed method comparable to the STOI measure and other SIP algorithms with similar properties as well as to existing monaural non-intrusive algorithms. When using neural networks for ASR, these typically include millions of parameters, and are trained using thousands of hours of speech material [58]. We are not aware of the existence of such large databases of measured intelligibility, and this consequently rules out the possibility of training similarly large neural networks for SIP. Instead, this work has been guided by the following hypothesis:

SIP is a simple problem in comparison with ASR and speech enhancement, and can therefore be solved by a comparatively smaller neural network.

Under this hypothesis, it should be possible to obtain good performance on the SIP task, without having to rely on massive quantities of training data. The hypothesis is based on the assumption that speech intelligibility can be effectively assessed by the presence or absence of a rather small number of spectro-temporal patterns in a signal. This assumption is motivated by the observation that existing non-intrusive SIP algorithms are able to predict intelligibility accurately across many conditions, by considering very simple modulation-domain structures [35], [36], [40]. In contrast, an ASR system must, in some form or another, store a detailed lexicon of all sounds present in speech, to allow for distinguishing and classifying these.

In order to make our work interpretable and comparable to existing non-intrusive and intrusive SIP algorithms, we have used a simple CNN structure, which is based on a small number of easy-to-visualize modulation features. In Sec. II, we provide a detailed description of the structure of this network. In Sec. III, we describe the speech database used for training, validation, and testing of the method. In Sec. IV, we evaluate the proposed measure, and compare it with existing non-intrusive and intrusive SIP algorithms. In Sec. V we discuss the implications of specific design choices in the proposed CNN architecture and the approach used for training it. Sec. VI concludes upon our findings.

II. A CNN for Intelligibility Prediction

In this section we describe the CNN architecture that we use to non-intrusively predict speech intelligibility. We specifically chose a CNN structure because 1) this allows for handling input signals of varying length, and 2) the resulting convolution kernels can be visually inspected, and clearly reveal the spectro-temporal features used by the network.

Since we consider non-intrusive intelligibility prediction, only a degraded input signal, $y(t)$, is available. This is assumed to be a recording including one or more spoken sentences, which may be degraded by noise, reverberation, non-linear distortion, or essentially any other factor. The goal is to predict the fraction of words that are understandable to a normal-hearing listener. By a word being "understandable", we mean that the listener is able to repeat it, after having heard it. In this study we define the ground truth intelligibility as an average obtained across multiple sentences and for multiple subjects but in the same listening condition. Thus, the output of the proposed method is a number in the range 0% to 100%.

Non-intrusive SIP algorithms are most often considered for real-time applications, where a clean signal is unavailable. The task of predicting intelligibility for a signal in real-time is illustrated in Fig. 1. This shows a noisy signal with a sentence placed in the middle. Underneath, illustrative values of true and estimated intelligibility levels are shown. The true intelligibility may have been obtained in a listening experiment where multiple sentences were presented to subjects in the same particular condition. Notably, it is a fixed value, which can only be considered valid within the speech-active region of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

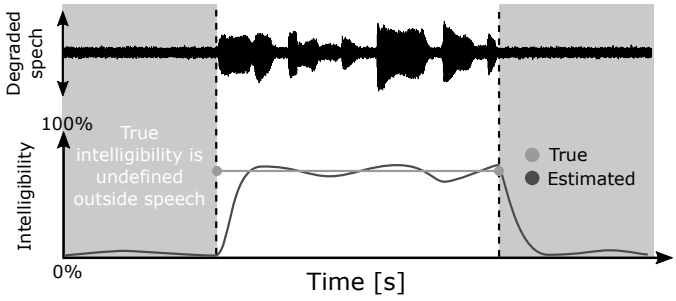

Fig. 1. An illustration of a real-time non-intrusive SIP algorithm.

the signal. Outside this region, speech intelligibility is poorly defined because no underlying speech is present. The estimated intelligibility is a time-varying signal based on a short history of the degraded signal (e.g. one or a few seconds). When speech is present, the estimate should ideally be close to the true intelligibility. When only noise is present we should expect the estimate to approach 0% as pure noise cannot be distinguished from noisy speech at a very low SNR. At the same time, it appears intuitively sensible that intelligibility is 0% when no speech is present.

For the present study we are interested in evaluating whether the proposed system is capable of correctly predicting measured intelligibility. Therefore, instead of generating time-varying estimates based on a short signal history, we allow the proposed system to integrate across a full signal consisting of multiple sentences to generate a single prediction of intelligibility. This prediction is compared with measured intelligibility (which is also based on multiple sentences). When doing this, it is important that the input signal does not contain long pauses where speech is absent, as speech intelligibility is undefined in these. In the evaluation of the proposed method, we therefore use an ideal Voice Activity Detector (VAD) to remove long stretches with no speech. It is important to stress that the use of an ideal VAD is necessary only to be able to *evaluate* the proposed method in a meaningful manner. Specifically, an ideal VAD is necessary to ensure that the evaluation of the system is only carried out on parts of the input signal where speech intelligibility is well defined and known. A VAD is not necessary for operation of the proposed method. For example, a VAD is not necessary to generate time-varying estimates as illustrated in Fig. 1 (we show examples of such estimates in Sec. IV-F). We refer to Sec. V for further discussion of the necessity and implications of evaluating the proposed system together with an ideal VAD.

The input signal, $y(t)$, is preprocessed before being presented to the aforementioned CNN. This is done to lower computational demands and to make the resulting network independent of the overall level of the input speech. Furthermore, the preprocessing steps serve to roughly model the frequency selectivity of the cochlea. Because the preprocessing also serves to lower the input signal dimensions, it may also be seen as a form of dimensionality reduction. The preprocessing consists of steps which are highly similar to ones carried out in the STOI measure [13]. These steps result in a considerably more compact signal representation, and the success of the STOI measure suggests that they do not remove information which is crucially important to speech intelligibility.

A. Preprocessing

The input signals are first resampled to 10 kHz and periods without speech are removed by use of an ideal VAD, i.e. a VAD which makes use of the underlying clean speech signal. As previously stated, the ideal VAD is necessary to meaningfully evaluate the performance of the trained network, but is not needed in real-world uses of the proposed system. The applied VAD consists of two steps: 1) both the clean and degraded signals are segmented into 256-sample Hann-windowed segments, with an overlap of 50% between consecutive frames, and 2) the degraded signal is resynthesized, using only frames where speech is present. We define speech-active frames as ones which contain clean speech energy in excess of -40 dB relative to the most energetic frame [13]. Furthermore, frames are also labelled as speech active if they do not belong to a sequence of at least one second where no frame has a clean speech energy in excess of -40 dB (i.e. only consecutive non-speech regions longer than one second are removed). Thereby, short pauses between words are categorized as speech active, so that the output of the VAD is still naturally sounding speech.

After resampling and removal of segments without speech, the signal is analyzed with a short-time Discrete Fourier Transformation (DFT). This is done in 256-sample Hann-windowed segments which are zero-padded to 512 samples. The DFT coefficient corresponding to the kth frequency bin and the mth time frame is denoted $\hat{y}_{k,m}$.

Envelopes are then extracted in $Q = 15$ one-third octave bands, across the M time frames in the signal [13]:

$$Y_{q,m} = \sqrt{\sum_{k=k_1(q)}^{k_2(q)} |\hat{y}_{k,m}|^2}, \tag{1}$$

for $m = 1, \ldots, M$ and $q = 1, \ldots, Q$, where q is the one-third octave band index, and $k_1(q)$ and $k_2(q)$ are, respectively, the lower and upper limits of the qth one-third octave band. The one-third octave bands have center frequencies spaced by one third octave, starting at 150 Hz.

The envelopes are mean- and variance normalized. We define the normalized envelope sample, $\bar{Y}_{q,m}$, by the two following steps:

$$\breve{Y}_{q,m} = Y_{q,m} - \frac{1}{N} \sum_{m'=m-N+1}^{m} Y_{q,m'}, \tag{2}$$

for $m = N, \ldots, M$, and:

$$\bar{Y}_{q,m} = \frac{\breve{Y}_{q,m}}{\sqrt{\frac{1}{N} \sum_{m'=m-N+1}^{m} \breve{Y}_{q,m'}^2}}, \tag{3}$$

for $m = 2N - 1, \ldots, M$, where $\breve{Y}_{q,m}$ is a zero-mean intermediate variable, and $\bar{Y}_{q,m}$ is the normalized envelope. We use $N = 30$ envelope samples (corresponding to 384 ms) to estimate the mean and variance. The resulting normalized envelopes are defined for $Q = 15$ one-third octave bands, and for $L = M - 2N + 2$ time windows.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

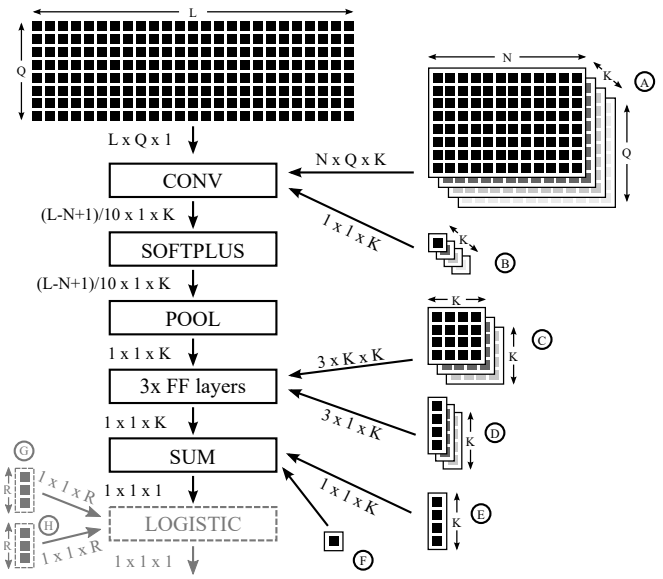

Fig. 2. A block diagram of the applied network architecture. Network weights, indicated by (A)–(H), are found by stochastic gradient descent. The output of the SUM-layer can be used as an index of intelligibility comparable to the SII or the STOI measure. By including the dataset-dependent LOGISTIC-layer, the network can be used to make direct predictions of intelligibility in percent.

B. Network Architecture

The preprocessed input signal can be represented as a matrix of $L \times Q$ real numbers. The aim is to make a prediction of the intelligibility of the input signal in the range 0–100%. The specific CNN architecture, used to do so, is illustrated in Fig. 2.

First, the preprocessed input signal is convolved with K kernels of dimension $N \times Q$ (marked (A) in Fig. 2). This convolution is carried out only for shifts where the kernel is entirely contained in the preprocessed signal. Furthermore, the convolution is carried out with a subsampling (or a stride) of 10, to limit computational demand. Thus, the output of this convolution is a tensor of dimension $\lfloor (L-N+1)/10 \times 1 \times K \rfloor$, where $\lfloor \cdot \rfloor$ is the floor operator. A kernel-dependent constant is added to this tensor (marked (B) in Fig. 2). The temporal length of the kernels has been chosen to $N = 30$ (or 384 ms), because a similar time scale was used with good results for both the STOI [13] and Non-Intrusive STOI (NI-STOI) [40] measures. The kernels were chosen to span all $Q = 15$ frequency bands, in order not to impose any constraints on the modelled structures across frequency. Furthermore, this design allows for easy visual inspection and interpretation of the kernels. The result of the convolution is transformed in an entry-wise manner with a "*softplus*" non-linearity [61]:

$$f(z) = \log(1 + e^z). \tag{4}$$

The softplus-function is used because it was found to lead to fast and stable convergence during training. Following this, pooling is carried out by averaging across the first (temporal) dimension of the signal, yielding an output vector of dimension $1 \times 1 \times K$. This vector is passed through three conventional Feed-Forward (FF) layers (weights marked (C) in Fig. 2), with K inputs and K outputs, each employing the softplus as activation functions. A constant-vector is added before each non-linearity (marked (D) in Fig. 2). A weighted summation of the the K outputs of the final FF layers is carried out (weights marked (E) and constant marked (F) in Fig. 2). The output of this summation (the SUM block on Fig. 2) is used as an index of intelligibility comparable to the SII or the STOI measure. In order to map this index to a direct prediction of intelligibility, we apply a logistic function exactly as for the STOI measure [13]:

$$f(x) = \frac{1}{1 + \exp(ax + b)}, \tag{5}$$

where x is the network output, and a and b are dataset dependent parameters. The resulting output is a scalar in the range 0 to 1, corresponding to 0–100% predicted intelligibility. When training the system with multiple datasets, we use separate values of a and b for each dataset (marked (G) and (H) in Fig 2, assuming that R datasets are used), but train these parameters exactly as the remainder of the network. That is, we perform training jointly across one instance of network weights together with one pair of logistic function parameters for each dataset used for training (a_r, b_r for $r = 1, \ldots, R$). See Sec. III for more details on the datasets used for training. The system was implemented using Theano [62].

C. Interpretation of the Architecture

The proposed CNN architecture can be related to the structure of existing SIP algorithms. The majority of existing SIP algorithms assume that additive contributions to intelligibility are supplied from different frequency bands [1], [4], [13] or modulation frequency bands [15], [28], [27]. Thus, contributions from several separate channels are computed and linearly combined, typically applying some numerical weighting of the importance of each channel [1], [63]. The contributions typically depend on either the SNR [1] or the correlation between clean and degraded envelopes [12], [13], [32], [11] within a band. For instance, the SII is computed as follows [4]:

$$\text{SII} = \sum_{i=1}^{n} I_i A_i, \tag{6}$$

where n is the number of frequency bands, I_i is the relative importance of the i'th band, and A_i is a measure of the speech fidelity in the ith band, which is, in turn, determined by transforming the SNR, in the ith band, with a compressive function. The resulting sum of contributions can then be transformed into a prediction of intelligibility (as a fraction of correctly understood words), by use of a mapping function with an output in the range 0 to 1 (e.g. a logistic function [13] or a cumulative normal distribution [27]).

In the CNN architecture proposed in this paper, the K outputs of the convolution stage react to the presence of different spectro-temporal patterns in the input signal. These excitations are non-linearly transformed and mixed in three FF layers. The outputs from the last FF layer are linearly combined using trained weights. The output of this stage can be considered as an index of intelligibility comparable to the SII or the STOI measure. This value may be transformed by a dataset-specific

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

logistic function to obtain a prediction of intelligibility in percent. The last stages of the CNN, specifically the SUM and LOGISTIC blocks in Fig. 2, are similar to existing SIP algorithms, in that they compute a weighted sum of contributions, which is then mapped onto the interval 0 to 1 with a mapping function. The last two steps of this process correspond closely to the computation of the SII, cf. (6). However, instead of combining the SNR-based values A_i, the CNN architecture combines the outputs of the FF network into an index. As for the SII (for example), this index may then be transformed into a prediction of intelligibility in percent by a logistic function.

The difference between the proposed CNN architecture and the SII is, then, that the proposed architecture non-intrusively estimates the K separate contributions, while the SII relies on the band-wise SNR, transformed by a compressive non-linearity. The proposed architecture does so by detecting the similarity of the input signals to a range of trained modulation templates (i.e. kernels), which may be indicators of speech, noise, or other factors that impact intelligibility. This in turn suggests that the combination of the convolutional stage and the FF network can be interpreted, at least structurally, as an SNR estimator, followed by a non-linear transformation similar to that used in the SII. An advantage of the proposed method is that visual inspection of the kernels can give an understanding of which features the network associates with speech (see Sec. IV-F).

D. Training

The system is trained in a supervised manner on a database of audio files of different lengths, each with an associated measured intelligibility (see Sec. III). Network weights (marked (A)–(H) in Fig. 2) were found using stochastic gradient descent [64], while optimizing for the cross entropy [64]. Each gradient step is computed from a minibatch assembled from 5 seconds of audio, picked from a uniformly distributed random location in the signal, from each of five randomly picked audio files from the training set, i.e a total of 5×5 seconds of audio, and 5 corresponding values of measured intelligibility between 0% and 100%. We consider one epoch to constitute a single use of each value of measured intelligibility. Consequently, one epoch makes use of only 5 seconds of audio from each associated audio file.

Training commences with a learning rate of 0.01. Performance is evaluated on both the training set and the validation set once every 50 epochs (see Sec. III for details on the training and validation sets). At this point, the learning rate is decreased by 15% if the current best training set performance has not been found within any of the last five performance evaluations. Training is stopped when the current best validation set performance has not been found within any of the last 20 performance evaluations. Final evaluation is performed using the weights which gave rise to the highest validation set performance.

The training process was regularized in two different ways. First, a regularization term, consisting of 10^{-5} times the sum of squares of all trained weights, was added to the objective function. Secondly, dropout, with a probability of 0.5 was used on the weights in the convolution stage. Dropout was not used in the remaining stages, as this was found to make the training process unstable. This is most likely because relatively narrow FF layers were used in practice (i.e. there were few nodes in each layer).

TABLE I
AN OVERVIEW OF THE FOUR LISTENING EXPERIMENTS.

ID	Collected by	# conditions
D_1	Kjems et al. [65]	114
D_2	Kjems et al. [65]	33
D_3	Jensen & Taal [11]	60
D_4	Studebaker & Sherbecoe [66]	318

III. DATA MATERIAL

To evaluate the proposed architecture, we assembled a dataset by combining data from a number of sources. This dataset was split into training, validation and testing sets in two different manners, as described in this section.

A. Sources of Data

We obtained measured intelligibility from four listening experiments described in the literature (see Table I):

- **D_1:** (Described in [65].) Intelligibility was measured for noisy sentences processed by Ideal Time Frequency Segregation (ITFS) (an idealized type of noise reduction where low-SNR DFT units are suppressed [65]). This was done for 1) four different noise types: Speech Shaped Noise (SSN), bottling factory hall noise, car noise, and café noise, 2) two different types of ideal binary masks, 3) eight different Relative Criterion (RC) values (parameter setting for the ITFS algorithms), and 4) three different SNR values. The measurements were carried out for 15 normal hearing subjects, using the Dantale II speech corpus [67], subjects responding verbally. In this work, we excluded the conditions with café noise, as this effectively consists of a single interfering talker. In conditions with a single interfering talker, a non-intrusive SIP algorithm must be supplied with additional information to correctly identify the target talker (as it could be any of the two talkers). We consider such conditions to be beyond the scope of this work.
- **D_2:** (Described in [65].) This experiment used the same speech material and noise sources as dataset D_1; however, ITFS processing was not applied. This was done for 15 normal hearing listeners with the Dantale II corpus, using an adaptive procedure for measuring the 20%– and 80% Speech Reception Thresholds (SRTs) [65][1] (subjects responding verbally). The speech intelligibility was estimated at 11 SNRs, uniformly spaced from -20 dB SNR to 5 dB SNR, by fitting a logistic function to the measured SRTs, and interpolating, as also done in [32]. The café noise conditions were removed from this dataset, for the same reason as for dataset D_1.

[1]The $X\%$ SRT is the SNR at which the subject is able to correctly repeat $X\%$ of presented words.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

- **D_3:** (Described in [11].) Intelligibility was measured for Dantale II sentences masked by ten different types of noise, including SSN, unintelligible babble (1, 2, and 6 talkers), sinusoidally intensity modulated SSN (2, 4, 8, and 16 Hz), "machine gun" noise, and "destroyer operations room" noise (the two last noise types are from the NOISEX corpus [68]). Intelligibility was measured at six SNRs, uniformly spaced by 3 dB, and centered around the experimenters rough estimate of the 50% SRT [11]. The experiment was carried out for 12 normal hearing listeners. The subjects responded by selecting words on a screen.
- **D_4:** (Described in [66].) Intelligibility was measured for high- and low-pass filtered spoken words masked by SSN [66]. This was done using recordings of the CID W-22 word lists [69]. Measurements were taken at 21 filter cutoff frequencies, from 112 Hz to 11 kHz, and 10 SNRs, from -10 dB to 8 dB. However, some combinations of cutoff frequency and SNR were left out of the study due to very low expected intelligibility (e.g. lowpass-filtered speech with a low cutoff frequency remains unintelligible at an SNR considerably above -10 dB, and intelligibility was therefore not measured for the lowest SNRs). Eight normal hearing listeners participated in the study, writing their responses on standardized paper forms.

The first three datasets were collected using the Danish Dantale II speech corpus [67] while D_4 was collected using the CID W-22 corpus [69]. It is worthwhile to notice that the Dantale II corpus includes only 50 unique words, while the CID W-22 corpus contains 200 unique words. By training on such small corpora, it is likely that the resulting CNN could be able to recognize individual words, or parts of words, and associate these with high levels of intelligibility. This is undesirable, as a truly non-intrusive SIP algorithm should not be dependent on a particular underlying speech corpus. To avoid this problem, we recreated the stimuli of all four experiments using another Danish speech corpus: Akustiske Databaser For Dansk (ADFD)[2]. This includes a wide variety of Danish sentences spoken by more than 600 individuals of both genders. In this way, it was possible to ensure that a broad corpora of sentences spoken by different, non-overlapping, sets of talkers were used for training, validation, and testing. For datasets D_1, D_2, and D_3, we were able to obtain the software used to generate the original stimuli, including the involved noise recordings and processing algorithms. This software was rerun using random sentences from the ADFD corpus instead of the Dantale II corpus as input. It was not possible to obtain any software or signals from the collection of dataset D_4. The signals were therefore recreated from the description given in [66]. Clean speech signals were generated by concatenating random ADFD sentences. Both highpass (HP) and lowpass (LP) filtering was carried out using 512th order linear phase Finite Impulse Response (FIR) filters designed using the windowing method. The clean speech was filtered before being mixed with noise. SSN was generated by filtering white noise to have the same long time spectrum as sentences from the ADFD corpus. The SNR was computed as the ratio of speech energy to noise energy *before* filtering the speech. Speech material corresponding to at least ten sentences per condition was generated for all four datasets. This resulted in slightly less than seven hours of audio in total (after voice activity detection).

As discussed in Sec. II-B, the proposed CNN architecture includes a separate logistic function for each dataset. However, due to the small size of dataset D_2, we use the same logistic function for datasets D_1 and D_2, because they were collected under nearly identical conditions.

The choices involved in assembling the above described dataset are further discussed in Sec. V.

B. Training, Validation, and Testing Sets

In total, the four datasets include measured intelligibility for 525 conditions (see Table I). To facilitate an evaluation of the proposed CNN architecture, these conditions were split into training, validation, and testing sets. The simplest way to do this would be to simply partition the data randomly into three subsets. However, before doing so, it is important to realize that varying levels of similarity exist between the conditions:

- For each combination of noise and processing, intelligibility was measured for multiple SNRs (counted as separate conditions). In other words, the exact same combination of noise and processing is present across multiple conditions.
- Each noise type may be present across multiple conditions differing only in the applied processing (and vice versa).

Thus, by splitting the conditions entirely randomly, it may be difficult to draw conclusions about what exactly the network has learnt. Specifically, prediction performance may depend to a large extent on whether similar conditions were included in the training and testing sets. Furthermore, the results may not be representative of the performance that could be obtained with entirely novel types of noise and processing.

To partly control for the above described issue, we investigate two different means of generating the training, validation, and testing subsets:

1) We perform the split such that, to the furthest extent possible, all noise/processing configurations are represented in all three subsets, but at different SNRs. In other words, we deliberately split the conditions such as to make the three subsets as similar as possible, in that they contain the same noise/processing conditions, only at different SNRs. This allows us to investigate the ability of the trained network to generalize to previously *unseen* SNRs in previously *seen* configurations of noise type and processing. A similar scheme was employed in [57].
2) We perform the split such that no noise/processing configuration is represented in more than one subset. This amounts to deliberately making the training and testing sets more dissimilar than they would have been if the split was carried out entirely at random. This allows

[2]See http://www.nb.no/sbfil/dok/nst_taledat_dk.pdf.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

us to investigate the ability of the trained network to generalize to *unseen* configurations of noise type and processing.

To perform the first type of split, the conditions are grouped according to the applied noise/processing combination (i.e. such that each group consists of conditions that differ only in SNR). Each group is split such that four of six conditions are placed in the training set and one of six conditions are placed each of the validation and testing sets. Whenever this distribution cannot be obtained exactly, the remaining data points are distributed via fair lottery[3]. This ensures that each combination of noise/processing is represented evenly across the training, validation and testing sets.

For the second splitting procedure we expect a considerable degree of variability in performance across different split realizations. This is expected because intelligibility may be more or less difficult to predict for the different noise/processing combinations (e.g. if a noise/processing combination, which is highly dissimilar from all others, is placed in the testing set, it may lead to low performance, and vice versa). To make sure that each condition is equally weighted in the analysis, we therefore use the second splitting procedure for k-fold cross-validation. To do so, the conditions are again grouped according to the applied noise/processing combination. Each group is then randomly assigned to one of six subsets. Four of these subsets are used for training, one for validation, and one for testing. By rotating which subsets are used for training, validation and testing, we obtain six different splits, such that every data point is included exactly once in a validation set and once in a testing set. This ensures that predictions are made for all conditions in the available dataset.

IV. Results

In this section we evaluate the proposed CNN, and compare it with four previously proposed non-intrusive and intrusive SIP algorithms. We separately evaluate performance for the two different approaches for generating training, validation, and testing subsets (described in Sec. III-B). We also evaluate the performance of the method for two additional datasets, which were not used for training, and for inputs of clean speech and pure noise. Furthermore, we investigate the impact of substituting one underlying clean speech corpus for another (as described in Sec. III-A). Lastly, we investigate kernels and time-varying predictions for a specific instance of a trained network.

A. Testing with Unseen SNRs

We first consider performance when training with all noise/-processing configurations, and testing with unseen SNRs (the first data splitting method described in Sec. III-B). Because of the relatively small size of the dataset, random factors in the splitting procedure may affect performance. We therefore drew ten realizations of the split, and evaluated performance for each of these.

[3]E.g. if eight data points are to be split, four are assigned to the training set, one is assigned to the validation set, one is assigned to the testing set, and each of the remaining two samples are independently assigned to the either the training, validation, or testing sets with probabilities $4/6$, $1/6$, and $1/6$, respectively.

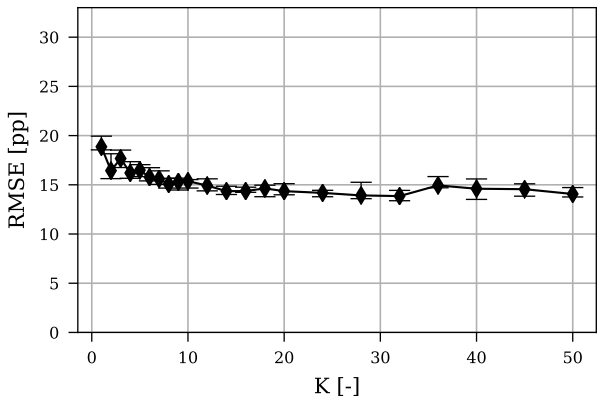

Fig. 3. The median test RMSE in percentage points (pp) vs K, when testing with unseen SNRs but previously seen noise/processing configurations. The error bars show 25th and 75th percentiles.

Fig. 3 shows the relationship between median prediction performance, in terms of Root-Mean-Square Error (RMSE), and the value of K (i.e. the number of convolution kernels). The median is computed across the performance of ten trained networks; one for each dataset split. The figure shows performance improving for increasing K, until around $K = 14$, whereafter performance improves only slowly. The small error bars indicate that performance is consistent across the different realizations of splits. In absolute terms, the best RMSE is slightly below 14 percentage points (pp).

Based on the results of Fig. 3, we find $K = 14$ to represent a good trade-off between performance and computational demand. We therefore use this value when testing with unseen noise/processing configurations, as described in the following section.

B. Testing with Unseen Noise/Processing Configurations

We now consider performance for noise/processing configurations that were not used for training. To do so, we split the dataset by the second method described in Sec. III-B. We generated ten sets of six splits, as described in Sec. III-B. One CNN was trained for each resulting split, yielding ten different predictions for each condition (i.e. 60 CNNs were trained in total). The medians of these predictions, for each condition, are plotted against measured intelligibility in Fig. 4a. This shows that accurate predictions are made for the majority of conditions. This, to a certain extent, suggests that the trained CNNs have learnt fundamental features that govern the intelligibility of speech. Especially the conditions of dataset D_4 consistently appear to be very accurately predicted. Substantial prediction performance is also seen for datasets D_1 and D_2, although intelligibility is considerably overestimated for a group of conditions from dataset D_1. Dataset D_3 is notably less accurately predicted. This dataset contains speech in different types of modulated or strongly fluctuating noise types. Intelligibility in such conditions has proved difficult to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

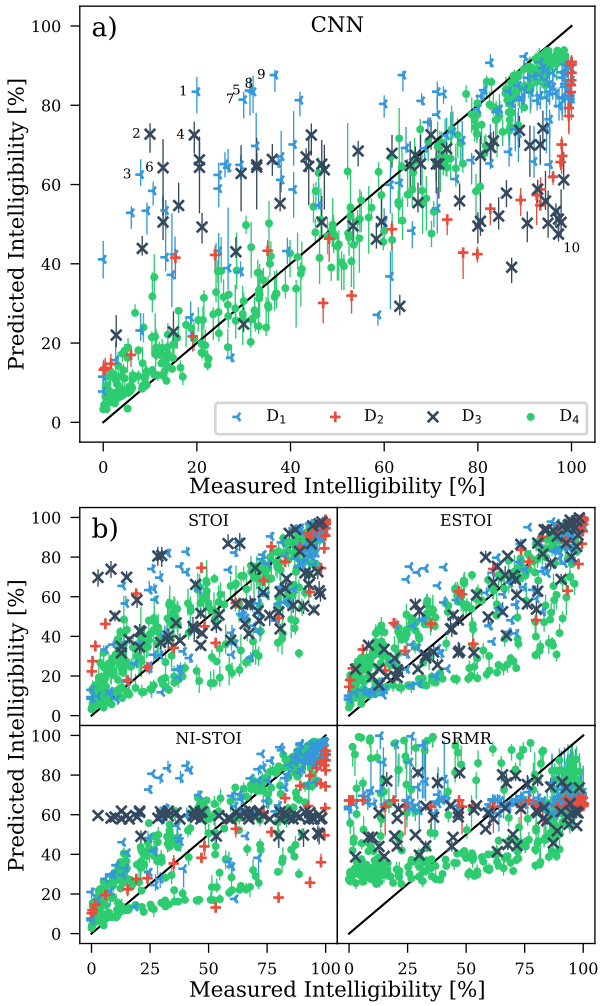

Fig. 4. The median of predictions for the proposed CNN, and for four other SIP algorithms, plotted against corresponding measured intelligibility. The error bars show the 25th and 75th percentiles of predictions. Colors/symbols indicate which dataset each condition belongs to. For the proposed method, the ten conditions with the largest absolute prediction errors are numbered in descending order. Descriptions of these are given in Table II.

predict by several existing SIP algorithms, including the SII and the STOI measure [6], [11].

The ten largest prediction errors on Fig. 4a are annotated and listed in Table II. Not surprisingly, these consist of conditions from datasets D_1 and D_3. These are conditions with either very low SNR or with noise types that have speech-like modulations. More surprisingly, a large prediction error occurs for speech in SSN at an SNR of -2 dB (No. 10 in Table II, dataset D_3). This error can most likely be attributed to a poorly fitted logistic function for dataset D_3, as prediction performance is generally low for this dataset.

Fig. 4b shows predictions by four other SIP algorithms: the intrusive STOI [13] and Extended STOI (ESTOI) [11] measures, as well as the non-intrusive NI-STOI measure [40] and the SRMR [35], [38]. To ensure a fair comparison, we preprocessed the input signals, for these algorithms, with exactly the same ideal VAD as used in preprocessing for

TABLE II
CONDITIONS WITH LARGEST ABSOLUTE PREDICTION ERRORS.

No.	Error		Description
1	63.4	pp	D_1 • Car cabin • SNR=–60dB • TBM RC=12.7dB
2	62.7	pp	D_3 • ICRA7: 6 spkr babble • SNR=–19dB
3	54.5	pp	D_1 • Bottl. fact. • SNR=–60dB • IBM RC=-34.9dB
4	53.1	pp	D_3 • ICRA7: 6 spkr babble • SNR=–16dB
5	52.4	pp	D_1 • Car cabin • SNR=–20.3dB • TBM RC=12.7dB
6	51.5	pp	D_3 • Mod. SSN f=2Hz • SNR=–27dB
7	51.4	pp	D_1 • Bottl. fact. • SNR=–60dB • TBM RC=-23.1dB
8	51.4	pp	D_1 • Car cabin • SNR=–23.0dB • TBM RC=12.7dB
9	50.9	pp	D_1 • Bottl. fact. • SNR=–60dB • IBM RC=-25.2dB
10	–49.7	pp	D_3 • ICRA1: SSN • SNR=–2dB

TABLE III
PERFORMANCE METRICS FOR THE FIVE SIP ALGORITHMS.

SIP algorithm	RMSE	Kendall's Tau
CNN	17.69 pp	0.667
STOI	18.94 pp	0.658
ESTOI	17.11 pp	0.692
NI-STOI	19.90 pp	0.629
SRMR	32.77 pp	0.281

the proposed architecture. The outputs of these predictors were transformed with a logistic function, (5), identical to the one used in the output of the proposed CNN architecture. The constants a and b were fitted to the training data[4], and predictions were carried out for the corresponding test set. One logistic function was fitted for each dataset (except for datasets D_1 and D_2 for which a single logistic function was fitted), as for the proposed CNN architecture. This was done for all 60 dataset splits, yielding ten different predictions for each point. Fig. 4b shows the medians of these.

The two intrusive SIP algorithms (the STOI and ESTOI measures) show consistently good prediction performance, with the exception of a group of conditions, for which the STOI measure severely underestimates intelligibility. These are mainly conditions from dataset D_3, which involve modulated or otherwise fluctuating interferers. The ESTOI measure was developed specifically to cope better with such conditions [11]. It is entirely expected that these algorithms perform favorably in comparison with the proposed one, as they have a considerable advantage, in having the clean signal available. This is not the case for the two non-intrusive SIP algorithms included in the analysis (the NI-STOI measure and the SRMR). The NI-STOI measure predicts the overall trend accurately, with the exception of dataset D_3. The SRMR correlates poorly with measured intelligibility in the studied conditions. It appears that this is partly caused by a number of conditions from datasets D_1 and D_4, for which intelligibility is substantially overestimated. This result has to be interpreted in light of the fact that the SRMR was initially developed with only reverberant speech in mind [35]. The developers of the SRMR have, however, later suggested its use for purposes beyond this [38], [25].

A notable feature of Fig. 4, is the fact that the measures in the STOI-family provide very consistent predictions across different training sets (i.e. the error bars are small), while

[4]The fitting was carried out using the `scipy.optimize.curve_fit`-function in SciPy [70].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

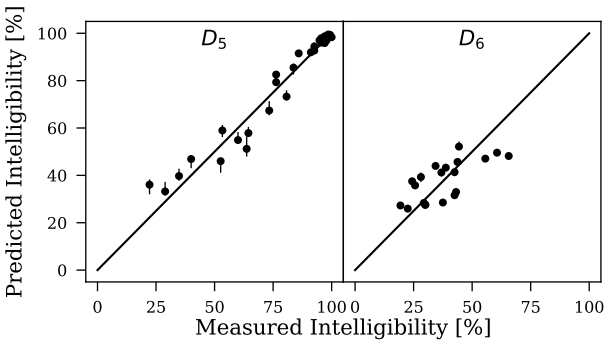

Fig. 5. Median predicted versus measured intelligibility for two datasets that were not used for training (D_5, D_6).

the CNN-based approach shows considerable variability. This is certainly a consequence of the fact that the STOI-family measures have only two fitted parameters (a and b in (5)), while the CNN-based approach has many more. However, the variations are rather small in comparison with the magnitude of prediction errors, and are therefore not likely to influence performance strongly. Even more consistent predictions could possibly be obtained with a larger and more representative dataset.

Table III lists the overall prediction performance of the five SIP algorithms in terms of RMSE and Kendall's Tau. Both performance measures show the proposed method to perform better than all but the ESTOI measure. This result should be considered in the light of the fact that the STOI and ESTOI measures gain a considerable advantage by having access to the clean speech signal. On the other hand, the CNN-based method is trained on conditions which, while not identical, may be similar to the ones used for testing.

C. Predictions for Entirely Unseen Data

The results presented above indicate that the proposed method may work well in unseen conditions. To explore this point further, we report results for two additional datasets, not used for training:

- **D_5:** (Described in [29].) Speech intelligibility was measured for sentences from the Dutch matrix sentence test [71] in SSN or car noise, unprocessed or processed by one of three different optimal energy redistribution algorithms at four different SNRs. The dataset has 32 conditions in total.
- **D_6:** (Described in [72].) Speech intelligibility was measured for sentences from the Dutch matrix sentence test [71] convolved with a room impulse response (T_{60} = 1s) and further degraded by additive SSN. The sentences were presented unprocessed or processed by one of four different optimal energy redistribution algorithms at four different SNRs. The dataset has 20 conditions in total.

Note that datasets D_5 and D_6 differ from the training set both in terms of speech material (Dutch sentences), processing type (energy redistribution), and distortion type (reverberation is included in D_6). Predictions were made directly from the stimuli, as presented to the subjects in the underlying listening experiments (i.e. no clean speech substitution was performed). Predictions were made using the output of the SUM-layer (see Fig. 2) as an index of intelligibility similar to the STOI measure or the SII. The indices were mapped to direct predictions of intelligibility by using logistic functions fitted separately to each of datasets D_5 and D_6. Predictions were made by all ten CNNs trained using unseen SNRs (Sec. IV-A) with $K = 14$. Results, shown in Fig. 5, indicate a high degree of prediction accuracy. This is supported by Table IV which shows RMSE and Kendall's Tau (note that Kendall's Tau is independent of the fact that the data points were transformed with a fitted logistic function, as this does not change the ordering of data points).

TABLE IV
MEAN PREDICTION PERFORMANCE FOR TWO LISTENING EXPERIMENTS WHICH WERE NOT USED IN TRAINING.

Dataset	RMSE	Kendall's Tau
D_5	5.57 pp	0.868
D_6	8.75 pp	0.546

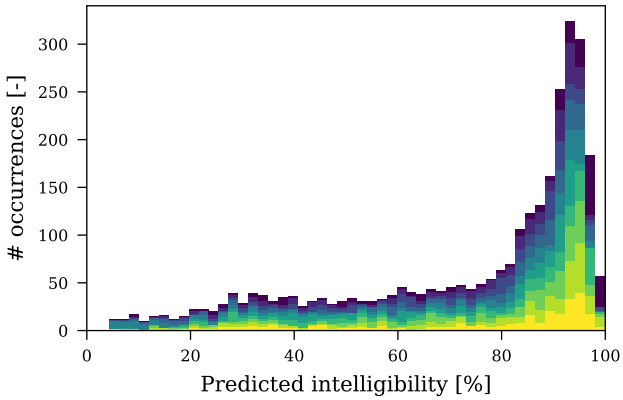

Fig. 6. A histogram of predicted intelligibility for each of 300 clean speech sentences from the ADFD corpus. Predictions were made with ten CNNs ($K = 14$). The bars are coloured according to the contributions of each CNN. Predictions were made with the logistic function corresponding to dataset D_4.

D. Predictions for Clean Speech and Noise

In order to further investigate the behaviour of the proposed method we show predictions obtained when using either clean speech or pure noise as input signals.

To make the results tangible, we used a logistic function to provide predictions in percent (i.e. the dashed block in Fig. 2) is included). We used the logistic function associated with dataset D_4, as this is the largest considered dataset and the one with the most robust results.

Results for clean speech were obtained by making predictions for 300 random ADFD sentences (sampled from the subset of the ADFD corpus used for testing). Predictions were made with each of the ten CNNs with $K = 14$ trained for the first type of data split. A histogram of the results are shown in Fig. 6. While a few sentences yield predictions substantially lower than 100%, the method clearly tends to predict high

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE V
PREDICTED INTELLIGIBILITY FOR VARIOUS NOISE RECORDINGS FROM THE NOISE-X DATABASE [68]. PREDICTIONS WERE MADE USING THE LOGISTIC FUNCTION ASSOCIATED WITH DATASET D_4.

Recording	Median prediction
White noise	1.2%
Pink noise	1.2%
HF channel	1.3%
Jet cockpit 1	1.3%
Jet cockpit 2	1.4%
Car interior	1.8%
F16 cockpit	1.8%
Destroyer engine room	2.0%
Factory floor 2	2.8%
Tank noise	2.9%
Military vehicle	3.3%
Destroyer operations room	7.5%
Factory floor 1	22.3%
Machine gun	38.6%
Speech babble	41.4%

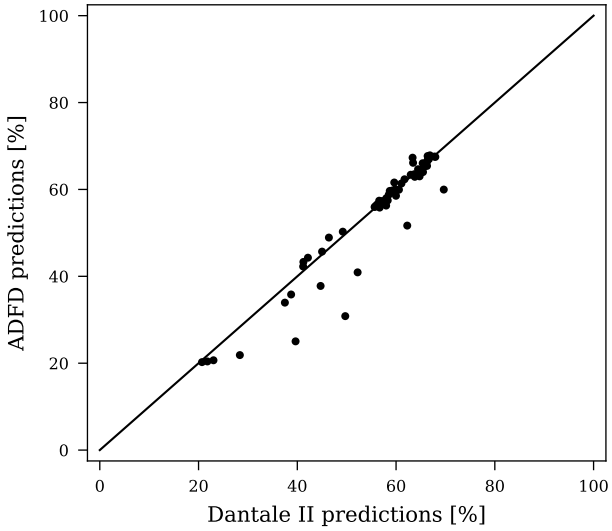

Fig. 7. Comparison of median predictions for dataset D_3 when using Dantale II and ADFD as underlying clean speech.

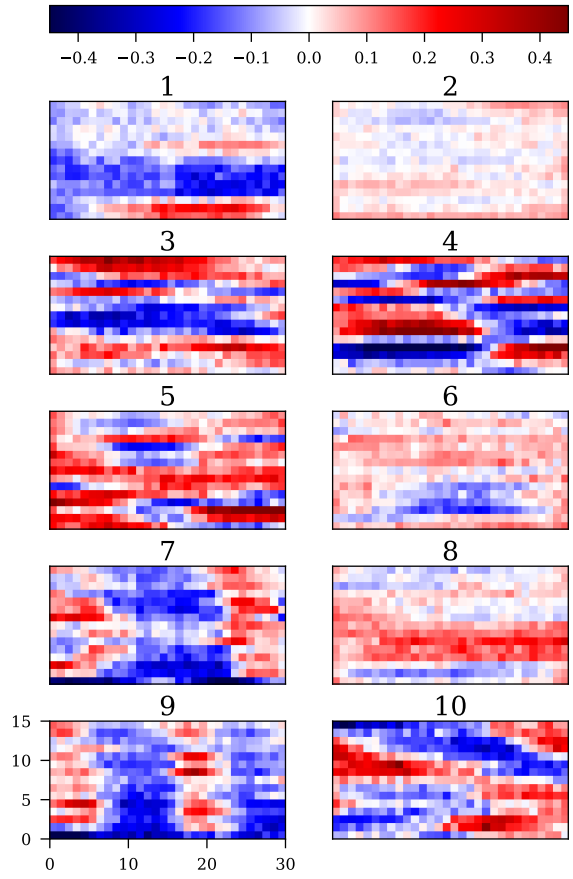

Fig. 8. Plots of the kernel weights for a CNN with $K = 10$, trained according to the procedure described in Sec. IV-A.

levels of intelligibility for clean speech. The median prediction is 85.3%.

Results for pure noise were obtained in a similar fashion, using noise recordings from the NOISEX corpus [68]. Median predictions for different noise types are listed in Table V. The results clearly show that predicted intelligibility increases for increasingly modulated noise sources. This shows that the trained CNNs associate high speech intelligibility with various forms of modulation.

E. Validation of Clean Speech Substitution

In this study, the underlying clean speech from the used datasets was substituted with clean speech from another corpus (the ADFD corpus). This was done to ensure that the proposed method was not over-fitted to the small speech corpora used in the original datasets. The assumption behind this is that predicted intelligibility does not change, when another speech corpus is substituted in place of the one used for collecting a dataset. To investigate the validity of this assumption, we made predictions for dataset D_3, using the original underlying clean speech (Dantale II) and compared these with the ADFD-based prediction (as shown in Fig. 4a). We did this analysis for dataset D_3 because it contains highly diverse noise types, and because it is appears to be the most challenging of the studied datasets, according to Fig. 4. The predictions with Dantale II were made in exactly the same manner as those using ADFD sentences. Median predictions are compared in Fig. 7. The figure verifies that for the vast majority of conditions, it makes only a small difference whether Dantale II or ADFD sentences are used as underlying clean speech material.

F. Interpretation of Trained Network

Because of the simple structure of the used CNN and the small values of K, it is possible to effectively visualize the operation of the resulting network. To do this, we selected one particular network with $K = 10$, trained according to the procedure applied in Sec. IV-A. In this section, we illustrate the operation of this network in further detail.

The kernel weights of this network are plotted in Fig. 8. Each kernel spans $Q = 15$ one-third octave bands, logarithmically spaced from 150 Hz to 3.8 KHz, and $N = 30$ time frames, corresponding to 384 ms. The convolutional part of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

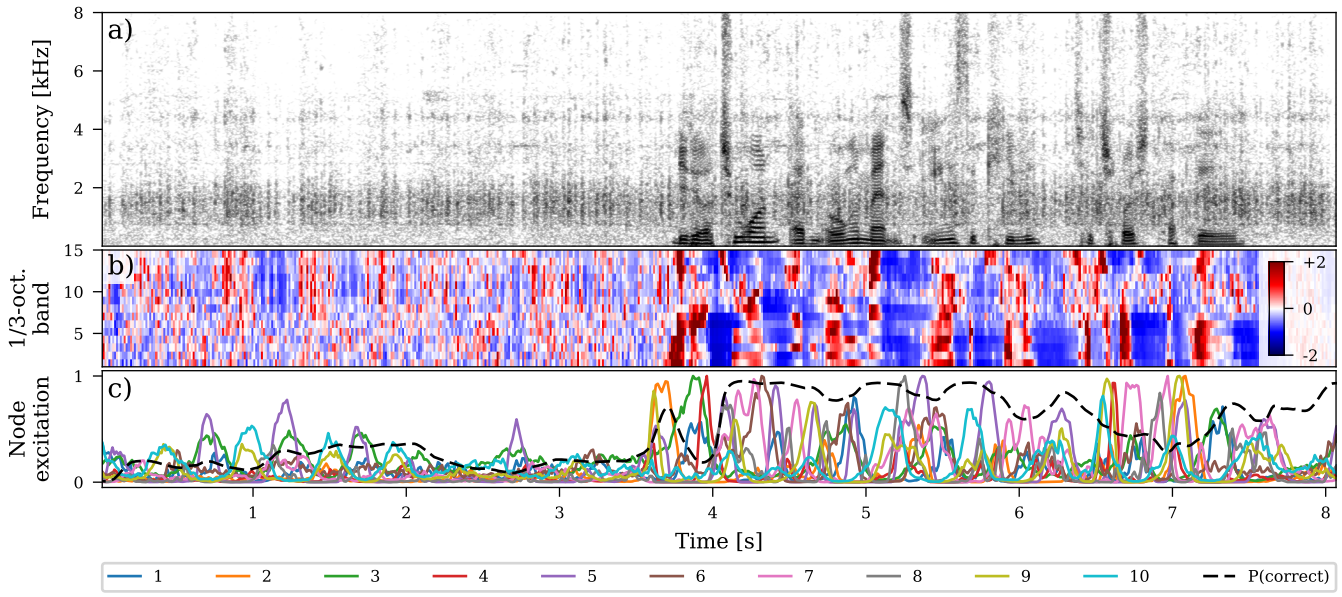

Fig. 9. a) A spectrogram of a signal consisting of about four seconds of bottling factory hall noise, followed by about four seconds of speech in bottling factory hall noise at an SNR of 10 dB. b) The signal representation used as input for the proposed CNN-based SIP algorithm. c) The corresponding instantaneous excitation of the ten kernels displayed in Fig. 8, as well as instantaneous predicted intelligibility based on one trailing second of signal. The excitation signals have been normalized to the interval 0 to 1.

network can be interpreted as if looking for segments of signal with a time-frequency pattern similar to the kernels. The output of the convolutional stage is passed through three FF layers. A high degree of excitation for a particular kernel, may be interpreted, by the later stages of the network, as an indication of either higher intelligibility or of lower intelligibility (e.g. a kernel can function as a detector of intelligible speech, but also as a detector of noise). Since the contribution of each kernel is determined through a non-linear mixing with other kernel excitation levels, it is not possible to generally map the contribution of each kernel to predicted intelligibility. The kernels seen in Fig. 8 clearly represent various types of spectro-temporal modulation structure. For instance, kernels 7 and 9 appear to encode temporal modulation at rates around 3 and 5 Hz, respectively. Kernel 7 could also be interpreted as a detector of short bursts (e.g. short voiced segments). Kernels 4, 5 and 10 appear to represent more complex temporal developments with the spectral distribution changing across time.

To further illustrate the operation of the network, Fig. 9 shows several internal properties of the network, for an input signal consisting of a short segment of bottling factory hall noise followed by a short segment of speech and bottling factory hall noise at +10 dB SNR. Fig. 9a shows a conventional spectrogram of the signal. Fig. 9b shows the signal representation used as input to the network (i.e. the representation given by (3)). Fig. 9c shows the ten outputs of the convolution stage (i.e. obtained by convolution with the kernels displayed in Fig. 8 followed by application of the softplus non-linearity), as well as the predicted intelligibility of the signal (obtained, for illustration, by using the logistic function associated with dataset D_4), computed based on one trailing second of signal. Note that this application of the proposed algorithm does not use an ideal VAD. The VAD mechanism is is implicitly taken care of by the network, and results in low predicted speech intelligibility in noise-only segments. From Fig. 9c, it is evident that the kernels are strongly excited by the speech part of the signal, and much less so by the non-speech part. The individual kernels are excited somewhat sparsely during the speech part of the signal, and each of them at different points in time. This could indicate that the training has caused the kernels to be excited by different features in speech signals. The predicted speech intelligibility, shown in Fig. 9c, is mostly low in the non-speech part of the signal, and becomes close to one in the speech part of the signal. This suggests that the network is able to distinguish speech from the noisy background. However, the strong fluctuations of the prediction also suggests that one second of audio may not be enough to accurately predict intelligibility (one could argue that speech intelligibility is not even a meaningful concept for such a short signal).

V. Discussion

In this section we further discuss some design decisions concerning the proposed CNN architecture and the procedure used for training it. Specifically we discuss 1) the use a an ideal VAD in the preprocessing step for an otherwise non-intrusive method, and 2) the joining of data across multiple listening experiments and the associated substitution of underlying clean speech material.

A. The Use of an Ideal VAD

The preprocessing steps carried out on the degraded speech inputs to the proposed CNN architecture (Sec. II-A) involve the use of an ideal VAD. The use of an ideal VAD together

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

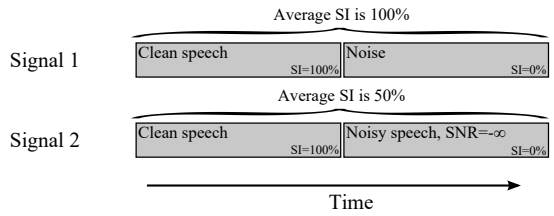

Fig. 10. Two hypothetical signals, each composed of two consecutive segments. The first signal consists of a token of clean speech (e.g. a sentence), and a token of pure noise. The second signal consists of a token of clean speech, and a token of noisy speech, presented at an SNR which makes it indistinguishable from pure noise. The speech intelligibility of the first signal is 100%, as only the first token contains intelligible speech, and the second token contains no speech. The average intelligibility of the second signal is only 50%, as the first token contains intelligible speech, while the second token contains unintelligible speech.

with an otherwise non-intrusive method may seem somewhat inconsistent. However, we argue that the use of an ideal VAD is necessary to ensure meaningful training and evaluation of the proposed method. When applying the proposed method in practice, no VAD is necessary.

To illustrate the necessity of an ideal VAD, two hypothetical signals are shown in Fig. 10. Each signal consists of two tokens of speech and/or noise. Signal 1 consists of one token of clean speech (e.g. a sentence), and a token of pure noise without speech. The first token is easily intelligible, and the second token contains no speech. Therefore all speech in Signal 1 is intelligible, i.e. the intelligibility is 100%. Signal 2 consists of one token of clean speech, followed by one token of noisy speech, presented at an SNR of $-\infty$ dB. In Signal 2, the first token of speech is intelligible, but the second one is not, i.e. the average intelligibility is 50%. However, there is no way, even in principle, to predict this difference from the degraded signals only, because signals 1 and 2 are, in fact, completely identical: both signals contain one token of clean speech, and one token of noise. The only difference is whether the second token is considered to contain an underlying, unintelligible, speech signal, or not. To meaningfully evaluate the performance of a non-intrusive SIP algorithm for these signals, one should remove the second half of Signal 1, which does not contain any underlying speech information. This makes it possible for the algorithm to realize that all speech in Signal 1 is clean, while the second half of Signal 2 is severely corrupted by noise. In this way, one evaluates how accurately the algorithm can predict speech intelligibility whenever speech is present. In exactly the same way, a listening experiments evaluates only how well a subject is able to understand speech when speech is present; not in the arbitrary pauses between sentences.

For practical uses of the proposed system it is unnecessary to use a VAD. In real-world applications, a system like the proposed one could be used to produce a time-varying estimate of intelligibility based on a few seconds of signal history (as shown in Fig. 9c). In such a use-case, the system will predict intelligibility when speech is present, and approach 0% when no speech is present. This aligns well with the intuitively expected behaviour for such a system, and does not rely on any type of VAD.

B. Joining Data Across Different Listening Experiments

Sec. III-A describes a process for replacing the underlying speech material, used in listening experiments, with speech from an alternative corpus. We assume that it is still meaningful to use such modified stimuli for predicting intelligibility. For this to be the case, the used CNN architecture should respond similarly to the original degraded speech and the degraded speech with substituted target speech. Because the employed architecture uses kernels with a duration of 384 ms, we assume it to be mainly sensitive to features on this time-scale, i.e. individual phonemes and transitions between these. The substitution should therefore be unproblematic, provided that the phonetic structures are similar across the original and substituted corpora. Since Dantale II and the ADFD corpora both contain common Danish sentences, we consider this to be a reasonable assumption for datasets D_1, D_2 and D_3. This is supported by evidence from Sec. IV-E. We were unable to obtain the used recordings of the CID W-22 word lists used in dataset D_4, and it is therefore more difficult to assess whether the substitution of speech corpora is justifiable in this case. Notably, the CID W-22 corpus contains English words rather than full Danish sentences. While we do not believe the difference between English and Danish to be highly important, it has previously been suggested that intelligibility is lower, when individual words are presented without a context [73]. Such general differences should, however, easily be accounted for by the use of separate logistic functions for the different datasets.

The above discussion suggests an even broader question: is it sensible to merge results from different listening experiments, carried out with different subjects, different equipment, in different conditions, using different speech corpora? The answer, of course, depends on the magnitude of differences between the listening experiments, and the required quality of the merged dataset. All four datasets, considered in this work, have been collected with normal hearing subjects, presented with diotic degraded speech via headphones. Unless any of the equipment used in collecting these datasets have strongly impacted the results, it is reasonable to assume that results can be compared across such studies. We believe the main difference to be the, already mentioned, difference between the used speech corpora. The typical method to account for such differences within the SIP community has been to apply separate mapping functions (e.g. logistic functions) [13], [1]. By fitting separate logistic functions to the different datasets, we have introduced a similar means to account for these differences.

VI. Conclusion

We have proposed a Convolutional Neural Network (CNN) architecture for use in Speech Intelligibility Prediction (SIP). The architecture is designed with a specific focus on being interpretable and structurally comparable to existing SIP algorithms. To evaluate the performance of the architecture, we collected a dataset of measured intelligibility by combining

the results of four listening experiments from the literature. The performance of the proposed method was shown to be similar to or higher than that of four existing intrusive and non-intrusive SIP algorithms. Furthermore, it was shown to account for the intelligibility of two datasets which were not used for training.

VII. Acknowledgments

The authors would like to thank Morten Kolbæk for providing a reorganized version of the "Akustiske Databaser for Dansk" corpus, as well as for several helpful discussions. We also wish to thank three very dedicated anonymous reviewers, whose suggestions have led to significant improvements of this work. The work was funded by the Oticon Foundation and the Danish Innovation Foundation.

References

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[2] H. Fletcher and J. C. Steinberg, "Articulation testing methods," *Bell System Technical Journal*, vol. 8, no. 4, pp. 806–854, Oct. 1929.

[3] J. B. Allen, "The articulation index is a shannon channel capacity," in *Auditory Signal Processing: Physiology, Psychoacoustics and Models*, D. Pressnitzer, A. de Cheveigne, S. McAdams, and L. Collet, Eds. Springer Verlag, 2004, pp. 314–320.

[4] "Methods for calculation of the speech intelligibility index," American National Standards Institute, New York, United States, Standard, 1997.

[5] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[6] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, Dec. 2006.

[7] M. Cooke, "A glimpsing model of speech perception," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, Mar. 2006.

[8] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[9] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 768–776, Aug. 2014.

[10] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 436–446, Jul. 2013.

[11] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[12] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[14] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, Jan. 1971.

[15] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[16] H. vom Hövel, "Zur bedeutung der übertragungseigenschaften des aussenohrs sowie des binauralen hörsystems bei gestörter sprachübertragung," Ph.D. dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 1984.

[17] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[18] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3678–3690, Sep. 2010.

[19] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.*, vol. 127, no. 1, pp. 387–399, Jan. 2010.

[20] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and validation of a binaural model for speech intelligibility in noise," *Hearing Research*, vol. 275, no. 1–2, pp. 96–104, May 2011.

[21] A. Chabot-Leclerc, E. N. MacDonald, and T. Dau, "Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain," *J. Acoust. Soc. Am.*, vol. 140, no. 1, pp. 192–205, Jul. 2016.

[22] T. H. Falk, S. Cosentino, J. Santos, D. Suelzle, and V. Parsa, "Non-intrusive objective speech quality and intelligibility prediction for hearing instruments in complex listening environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, Canada: IEEE, May 2013.

[23] J. F. Santos and T. H. Falk, "Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2197–2206, Dec. 2014.

[24] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, Jul. 2014.

[25] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[26] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoust. Soc. Am.*, vol. 100, no. 3, pp. 1703–1716, Sep. 1996.

[27] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, Sep. 2011.

[28] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, pp. 331–348, 2003.

[29] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 303–307, Mar. 2015.

[30] S. Stadler, A. Leijon, and B. Hagerman, "An information theoretic approach to predict speech intelligibility for listeners with normal and impaired hearing," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.

[31] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, Jan. 2014.

[32] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, Feb. 2014.

[33] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[34] P. Loizou and J. Ma, "Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms," *J. Acoust. Soc. Am.*, vol. 130, no. 2, pp. 986–995, Aug. 2011.

[35] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[36] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, pp. 311–314, Dec. 2013.

[37] S. Cosentino, T. Marquardt, D. McAlpine, J. F. Culling, and T. H. Falk, "A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals," *J. Acoust. Soc. Am.*, vol. 135, no. 2, pp. 796–807, Feb. 2014.

[38] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. Juan les Pins, France: IEEE, Sep. 2014, pp. 55–59.

[39] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *The European Signal Processing Conference (EUSIPCO)*. Budapest, Hungary: EURASIP, Aug. 2016, pp. 1358–1362.

[40] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, United States: IEEE, Mar. 2017, pp. 5085–5089.

[41] C. Sørensen, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, United States: IEEE, Mar. 2017, pp. 386–390.

[42] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1363–1374, Sep. 2014.

[43] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, pp. 1016–1025, 2015.

[44] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015.

[45] ——, "Predicting the intelligibility of noisy and non-linearly processed binaural speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1908–1920, 2016.

[46] L. Lightburn and M. Brookes, "A weighted STOI intelligibility metric based on mutual information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, Mar. 2016, pp. 5365–5369.

[47] ——, "SOBM - a binary mask for noisy speech that optimizes an objective intelligibility measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Melbourne, Australia: IEEE, Sep. 2015.

[48] M. R. Schädler, A. Warzybok, S. Hochmuth, and B. Kollmeier, "Matrix sentence intelligibility prediction using an automatic speech recognition system," *International Journal of Audiology*, vol. 54, pp. 1–8, Sep. 2015.

[49] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech & Language*, vol. 48, pp. 51–66, 2017.

[50] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *The European Signal Processing Conference (EUSIPCO)*. Aalborg, Denmark: EURASIP, Aug. 2010, pp. 1899–1903.

[51] D. Sharma, Y. Wang, and P. A. Naylor, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Communication*, vol. 80, pp. 84–94, Apr. 2016.

[52] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, Mar. 2016, pp. 624–628.

[53] J. Xupeng and L. Dongmei, "A data-driven speech intelligibility assessment method using sum-sorted spectrogram feature," in *IEEE International Conference on Signal Processing (ICSP)*. Chengdu, China: IEEE, Nov. 2016, pp. 541–544.

[54] M. I. Mandel, "Learning an intelligibility map of individual utterances," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, United States: IEEE, Oct. 2013.

[55] ——, "Measuring time-frequency importance functions of speech with bubble noise," *J. Acoust. Soc. Am.*, vol. 140, no. 4, pp. 2542–2553, Oct. 2016.

[56] A. H. Andersen, E. Schoenmaker, and S. van de Par, "Speech intelligibility prediction as a classification problem," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. Vietri sul Mare, Italy: IEEE, Sep. 2016.

[57] A. Alghamdi and W.-Y. Chan, "Single-ended intelligibility prediction of noisy speech based on auditory features," in *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. Windsor, Canada: IEEE, Apr. 2017, pp. 386–390.

[58] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Oct. 2012.

[59] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3029–3038, Oct. 2013.

[60] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 153–167, 2017.

[61] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, "Incorporating second-order functional knowledge for better option pricing," in *Neural Information Processing Systems (NIPS)*. Vancouver, Canada: MIT Press, Oct. 2001, pp. 1369–1376.

[62] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

[63] J. M. Kates, "Improved estimation of frequency importance functions," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. EL459–EL464, Nov. 2013.

[64] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2017.

[65] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

[66] G. A. Studebaker and R. L. Sherbecoe, "Frequency-importance and transfer functions for recorded CID W-22 word lists," *Journal of Speech and Hearing Research*, vol. 34, pp. 427–438, Apr. 1991.

[67] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[68] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[69] I. J. Hirsh, H. Davis, S. R. Silverman, E. G. Reynolds, E. Eldert, and R. W. Benson, "Development of materials for speech audiometry," *The Journal of Speech and Hearing Disorders*, vol. 17, no. 3, pp. 321–337, Sep. 1952.

[70] E. Jones, T. Oliphant, P. Peterson, and others, "SciPy: Open source scientific tools for Python," Retrieved from `https://www.scipy.org` on 28/07-2017, 2001–. [Online]. Available: http://www.scipy.org/

[71] R. Houben, J. Koopman, H. Luts, K. C. Wagener, A. van Wieringen, H. Verschuure, and W. A. Dreschler, "Development of a Dutch matrix sentence test to assess speech intelligibility in noise," *International Journal of Audiology*, vol. 53, no. 10, pp. 760–763, Oct. 2014.

[72] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time sii," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 5, pp. 851–862, 2015.

[73] G. A. Miller, G. A. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test materials," *Journal of Experimental Psychology*, vol. 41, no. 5, pp. 329–335, May 1951.

Asger Heidemann Andersen received the B.Sc. degree in electronics & IT, the M.Sc. degree in wireless communication (*cum laude*), and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 2012, 2014, and 2017 respectively. While pursuing the Ph.D. degree he was employed at Oticon A/S and associated with the signal and information processing section at Aalborg University. He is currently employed at Oticon A/S as an audiology and DSP developer. His main research interests are prediction and measurement of speech intelligibility and enhancement with applications to hearing aids.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2847459, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Jan Mark de Haan received the M.Sc. degree in Electrical Engineering and the Ph.D. degree in Applied Signal Processing from Blekinge Institute of Technology, Karlskrona, Sweden, in 1998 and 2004, respectively. From 1999 to 2004 he was a Ph.D. student with the Department of Applied Signal Processing, Blekinge Institute of Technology. In 2003 he was a visiting researcher at the Western Australia Telecommunication Research Institute, Perth, Australia. Since 2004 he is employed at Oticon A/S, Copenhagen, Danmark. His main interest are in acoustic signal processing and signal processing applications in hearing aids.

Zheng-Hua Tan (M'00–SM'06) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999. He is a Professor in the Department of Electronic Systems at Aalborg University, Aalborg, Denmark. He is also a Co-Head of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge, USA, an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University, and a postdoctoral fellow in the Department of Computer Science at Korea Advanced Institute of Science and Technology, Daejeon, Korea. His research interests include machine learning, deep learning, pattern recognition, speech and speaker recognition, noise-robust speech processing, multimodal signal processing, and social robotics. He has authored/co-authored more than 170 publications in refereed journals and conference proceedings. He is a member of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee (MLSP TC). He has served as an Editorial Board Member/Associate Editor for Computer Speech and Language, Digital Signal Processing, and Computers and Electrical Engineering. He was a Lead Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and a Guest Editor of several journals including Neurocomputing. He is the General Chair for IEEE MLSP 2018 and was a Technical Program Co-Chair for IEEE Workshop on Spoken Language Technology (SLT 2016).

Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is a Senior Researcher with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is a Professor with the Section for Signal and Information Processing (SIP), Department of Electronic Systems, at Aalborg University. He is also a co-founder of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.